Universiteit
Leiden
Governance and Global Affairs

Charting Reproducibility and Performance

LLMs in Multilingual Toxic Speech Detection

Bastián González-Bustamante
Leiden University
✉ b.a.gonzalez.bustamante@fgga.leidenuniv.nl

Presentation at the LLM Pre-Conference Workshop, European Political Science Association (EPSA)
Universidad Carlos III de Madrid, Spain, June 25, 2025

Introduction

The Challenge of Online Toxicity

Undermining digital spaces. Online toxicity depresses participation, amplifies polarisation, and erodes trust in public (online) deliberation.

Contextual complexity. What one culture judges as banter, another may deem a slur. Power asymmetries further complicate judgements.

Evolving challenge. Early moderation technologies struggled with sarcasm and evolving slang.

Artwork by DALL·E 3 model

Research Questions

Artwork by DALL·E 3 model

LLLMs can now help to automate the detection and mitigation of abusive content. However, their performance remains uneven: accuracy fluctuates with language resources, model training processes, architecture, and deployment settings.

RQ1. Can open-source models match closed leaders beyond English and high-resource languages?

RQ2. Do chain-of-thought (CoT) routines help or harm performance in toxicity classification?

- Straightforward binary judgement

Empirical Expectations

Research Hypotheses

High-Resource Advantage Hypothesis

LLMs zero-shot toxicity classification will achieve higher performance for high-resource languages (i.e., English, German and Spanish).

Low-Resource $\times$ Openness Penalty Hypothesis

Proprietary, closed-source LLMs will outperform open-source LLMs in zero-shot toxicity classification for low resource languages (i.e., Arabic, Chinese, Hindi, and Russian).

Research Hypotheses

Overthinking Penalty Hypothesis

LLMs explicitly optimised for chain-of-thought (CoT) reasoning will achieve worse performance than general-purpose LLMs in zero-shot toxicity classification.

Small Models Equivalence Hypothesis

SLMs optimised for efficiency, fast inference, and low memory will match the performance of general-purpose LLMs in zero-shot toxicity classification.

› Temperature is used as a control in our meta-analysis and for some experiments with inter-coder measures.

Methods

Ground-Truth Data

Arabic
5000 tweets manually annotated

Chinese
5000 messages for toxic detection

English
5000 Wikipedia comments

German
5000 Twitter and Facebook comments

Hindi
5000 Twitter and Facebook comments

Russian
5000 comments on social network OK

Spanish
5000 messages for toxic detection

Note. We split the samples in a proportion of 70/15/15 for training, validation, and testing in case of fine-tuning jobs. The samples correspond to ground-truth data prepared for CLEF TextDetox (2024).

LLMs for Zero-Shot Classification

Our **prompt strategy** was based on the core definitions of Perspective and Google for toxicity with the following system message:

System message

Classify the category of the comment as either **TOXIC** or **NONTOXIC**. **TOXIC:** Rude, disrespectful, or unreasonable comments that are likely to make someone leave the discussion or stop sharing their perspective. **NONTOXIC:** Civil or nice comments that are unlikely to discourage conversation.

Along with providing texts of our balanced sample, we also listed the categories for the task as follows: **"Respond with only the category (TOXIC or NONTOXIC). Do not provide any additional analysis or explanation"**.

Zero-Shot In-Context Classification

109 LLMs

Run **807 times** under different conditions (e.g., parameters, API/local, temperature, datasets/language) for

(1) **error-rate analysis**
(2) **meta-analysis**

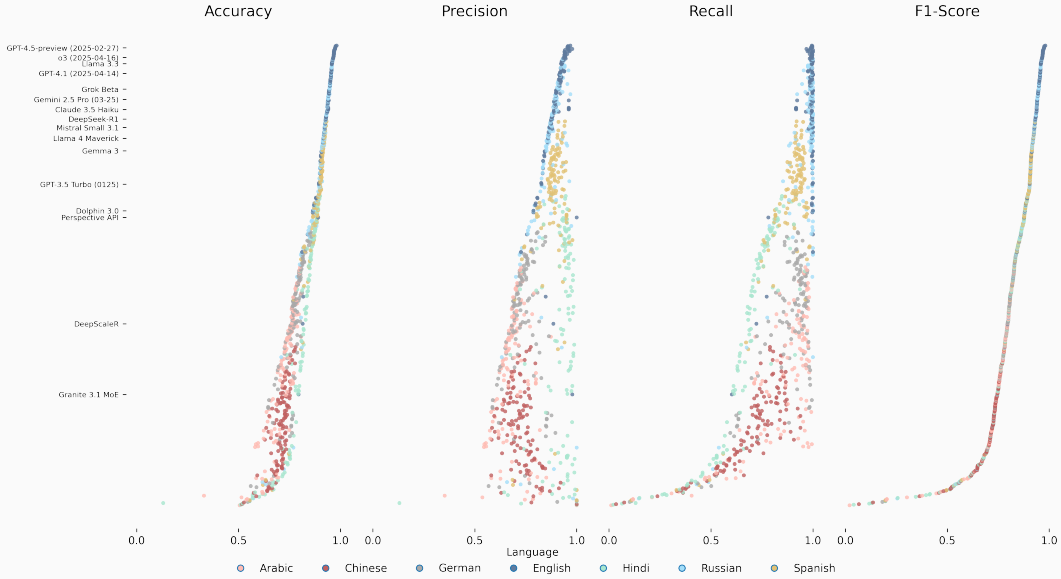
SOTA closed-source LLMs

o3, o4-mini, GPT-4.1, GPT-4.5-preview, Gemini 2.5, Grok 3 Beta, Claude 4 Opus

SOTA open-source LLMs

Llama 4 Maverick (400B) and Scout (107B), Mistral 3.1 (24B), Llama 3.3 (70B), DeepSeek-R1 (671B), DeepSeek-V3 (671B)

Results


Accuracy
Precision
Recall
F1-Score
GPT-4.5-preview (2025-02-27)
o3 (2025-04-16)
Llama 3.3
GPT 4.1 (2025-04-14)
Grok Beta
Gemini 2.5 Pro (03-25)
Claude 3.5 Haiku
DeepSeek-R1
Mistral Small 3.1
Llama 4 Maverick
Gemma 3
GPT 3.5 Turbo (0125)
Dolphin 3.0
Perspective API
DeepScaleR
Granite 3.1 MoE
0.0
0.5
1.0
Language
Arabic
Chinese
German
English
Hindi
Russian
Spanish

Top Performing Models

Arabic

0.83 F1-Score

o3 · o1 · GPT-4.5

Chinese

0.78 F1-Score

GPT-4.1 · GPT-4o · Grok 3

English

0.98 F1-Score

Granite 3.2 · OLMo 2 · GPT-4.5

German

0.86 F1-Score

GPT-4.1 · o3 · GPT-4.5

Hindi

0.89 F1-Score

Mistral Saba · Mistral Medium 3 · Gemma 2

Russian

0.96 F1-Score

o3 · GPT-4.1 nano · Claude 4 Opus

Spanish

0.93 F1-Score

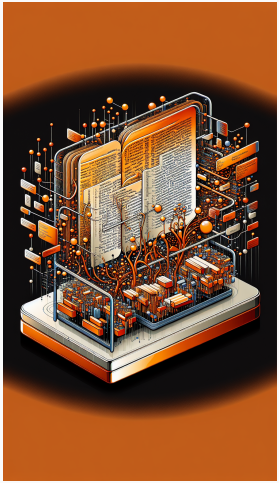
GPT-4.5 · o3 · GPT-4.1

Comparison to Perspective API

Perspective API performance. Competitive accuracy but abysmally low recall scores, resulting in low F1-scores across languages.

Performance gap. Even weakest LLMs outperform Perspective by about 0.20 points in F1-score for Arabic and Chinese.

Technological shift. Highlights transition from classical transformer pipelines to modern LLMs for toxicity classification.

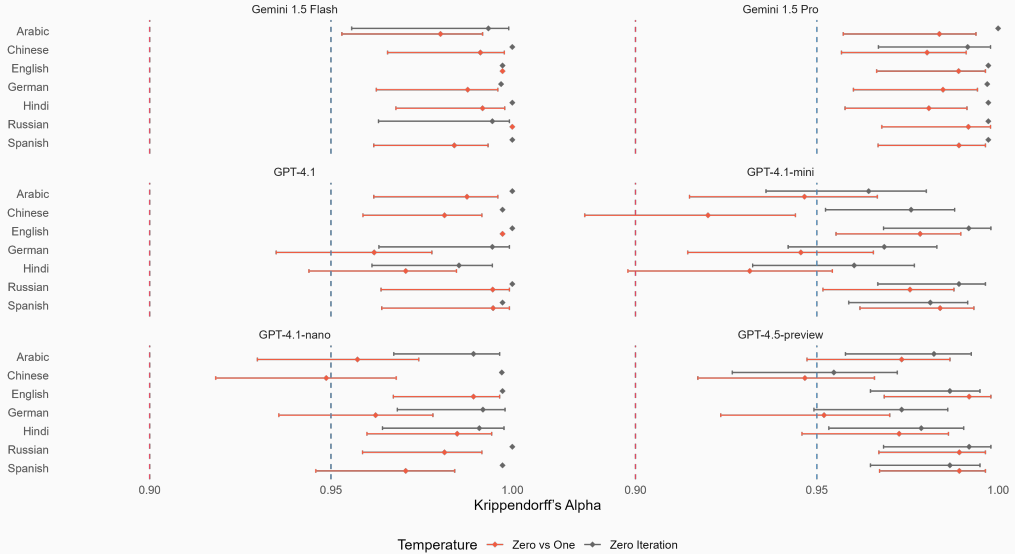


Artwork by DALL·E 3 model

Meta-Analysis

	Model I	Model II	Model III	Model IV	Model V
High-resource languages	0.711*** (0.059)	0.580*** (0.096)	0.580*** (0.096)	0.580*** (0.092)	0.580*** (0.092)
Open-source		−0.446*** (0.079)	−0.444*** (0.079)	−0.275*** (0.079)	−0.157* (0.089)
LRMs			0.028 (0.077)	0.107 (0.075)	0.073 (0.077)
SLMs				−0.644*** (0.080)	−0.530*** (0.087)
High-resource × Open-source		0.208* (0.120)	0.208* (0.120)	0.211* (0.116)	0.211* (0.115)
Constant	1.364*** (0.039)	1.646*** (0.063)	1.640*** (0.064)	1.624*** (0.062)	1.393** (0.132)
Parameters	No	No	No	No	Yes
Temperature	No	No	No	No	Yes
N	807	807	807	807	807
τ	0.826	0.807	0.808	0.776	0.772
R^2	0.153	0.192	0.191	0.252	0.260

Temperature Experiments



Takeaways

Key Findings

+7.7%

Resource Advantage

High-resource languages enjoy a 7.7 points F1 advantage

−9.8%

Size Penalty

Small models ($\leq$7B) fall 9.8 points short of larger LLMs baseline

99%

Annotation Reliability

Modern private LLMs exhibit outstanding reliability even w/temperature variations

We did not find 🔒 openness or 💡 overthinking penalties

Evidence-Based Recommendations

Skip reasoning models. LRMs exhibit longer latencies and higher costs but yield no performance gain for toxicity classification.

Small models for edge cases. Use smaller models only when memory, speed, or local deployment is necessary, accepting the performance penalty.

Set temperature to zero for full reproducibility. Maintain stable temperature and other parameters (i.e. repeat penalty, nucleus and topk sampling).

Performance vs reproducibility. Switching from deterministic to stochastic decoding rarely alters more than 1% of labels on average. The gain in performance could be expensive, depending on the model (API costs), but the reliability is better than that of human annotation ($\alpha \geq 0.90$).

Thank you very much!

Do you have any questions?

b.a.gonzalez.bustamante@fgga.leidenuniv.nl

bgonzalezbustamante.com

Presentation compiled with LaTeX and some

Download the latest version from GitHub